BACHELOR'S THESIS

# Degree in Mathematics

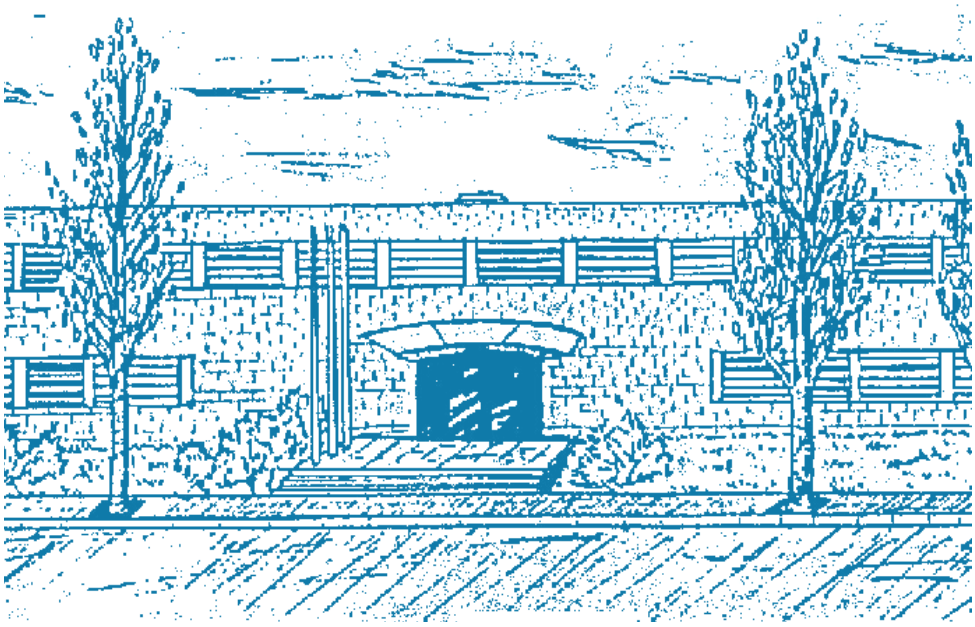
**Title:** Principal Component Analysis in running trainings

**Author:** Gerard Romagós Vilà

**Advisor:** Antonio Rodríguez-Ferran // Alberto García González

**Department:** Department of Civil and Environmental Engineering - LaCàN

**Academic year:** 2019-2020



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Degree in Mathematics

Bachelor's Degree Thesis

# Principal Component Analysis in running trainings

**Gerard Romagós Vilà**

Supervised by

Antonio Rodríguez-Ferran
Professor at UPC
LaCàN Group

Alberto García González
Professor at UPC
LaCàN Group

June, 2020

*Als meus tutors Antonio i Alberto, per haver-me guiat en aquest projecte,*
*a la meva família i amics, pel suport durant tota aquesta etapa,*
*i a la Marina, pels ànims i forces que m'ha donat en cada moment.*

# Abstract

Running training plans have become a trend between professional and amateur runners in recent years. In fact, everyone who trains for a race follows a training plan, even if it is done by himself/herself. The elabration of all of this training plans has been possible thanks to smartwatches that collect all kind of information of running trainings such as distance, pace, heart rate or elevation gain. However, some of the training plans have been elaborated without a preliminary study on what do they need to focus on to help the runner to be faster.

For this reason, in this document we are going to analyse running trainings in order to find which are those parameters that a training plan needs to improve to allow the runner to be faster. We will define a variable as a description of a training by taking its most general parameters and we will apply the Principal Component Analysis (PCA) method, due to it will let us discard those irrelevant parameters. We will see that they are not enough to classify a running training, so we will need to add more specific parameters to our variable and apply again PCA. This time, we will determine that a running training is defined by the minimum pace. This means that training and improving this parameter is going to be the the key for the training plan to turn the runner faster, because it will allow the athlete to be more comfortable in higher velocities.

Moreover, we are going to observe, thanks to PCA, that a training plan for a 5 or a 10 kilometers race will be the same because Distance is not going to be a relevant parameter. Also, we will notice that heart rate loses pace dependance when the runner runs faster than his/her threshold pace value.

Finally, as an application of our results, we are going to use the previous information to check that our particular training plan is helping us to improve our minimum pace and, consequently, is allowing us to be faster.

### Keywords

Running training plans, Smartwatches, Running training, Parameters, Principal Component Analysis (PCA), Pace, Distance, Heart rate, Threshold pace value.

# Contents

# Chapter 1

# Introduction

## 1.1 Background

In recent years we have seen that all of the sports World Records have been broken, with stratospheric performances that all of us thought they were inhuman few years before. Records like the sub-2-hour marathon in Vienna in 2019 [4] or the sub-8-hour ironman in the Ironman World Championship in Hawaii in 2018 and 2019 [6] show the improvement of all the athletes in sports.

The life of a professional athlete is based on training a lot of hours every day doing your best at every moment, resting the hours the body needs and eating a balanced diet. In short, they take care of their body in order to be able to continue training and improving every day. But in the lasts years, all of these things are not being enough to be a professional athlete. And it is at this point where the data analysis becomes the key to be the best.

Data analysis is used in athletics in many ways such as technology to have better materials that help the athlete to be faster, nutrition to control what is the athlete eating every day, in training planning to optimize workouts or in the same trainings to know in real time how is the training session going. In all of these examples, data analysis has caused a revolution and a need to have everything controled and analysed.

In the case of the real time information and the optimization of training plans, collecting all this information has been possible thanks to devices like smartwatches, heart rate sensors, GPS, power meters, etc. The integration of all of these sensors in the same device offers to the athlete the ability to know immediately how is he training, so he can adjust himself/herself his performance immmediately in order to follow as close as possible the theoretical training.

Focusing on running trainings, these devices create at the end of the training many graphics and pictures where the athlete can see how he has trained, with a lot of detailed information that will be used to improve in next trainings. For example, the following images show the route of a training and three graphics containing the elevation, the instant average pace and the instant and average heart rate at every time of a training.

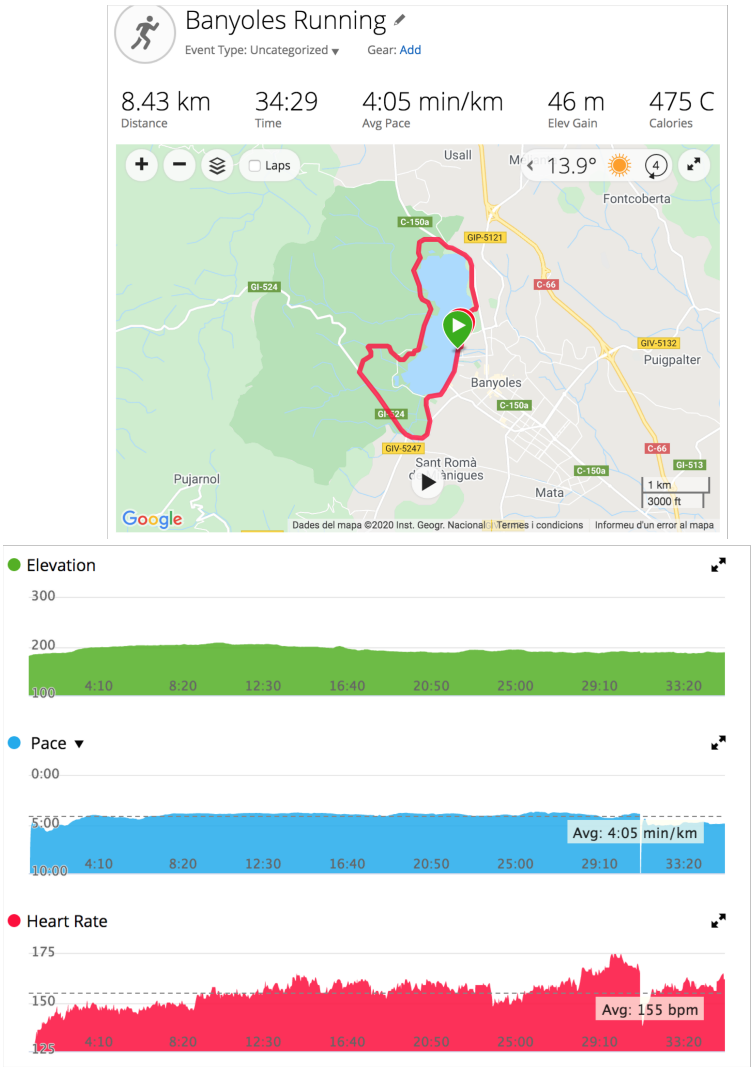



Figure 1.1: a) Route of a training. b) Graphics of the information of the training collected by the device.

All of this analysis is done with the purpose of being better and better. For runners, when they have a race, their objective is cross the finish line spending the least time possible. This is the reason why data analysis has made a change, because it allows the athletes to have controled much more aspects than before and train those parts of the race that they need.

## 1.2 Motivation

Millions of people arround the world like to do sport in their free time. I consider myself one of these people, specially with swimming, cycling and running. I am a person who needs to practice sport and I have found in triathlon the way to feel free and disconnect from daily preocupations. For this reason, and as a potential mathematician, I have felt the need to join my hobbies and my studies and analyse running trainings as an application of maths in sport.

## 1.3 Objectives

Thanks to the huge amount of data collected by the devices, the creation of training plans has grown exponentially. One of the aims of this project is to determine the most important parameters of a running training session, so that training plans will know where to focus on to improve and allow the athlete to get better results. At the same time, this information can be used by people to choose an appropiate training plan and discard the others.

Moreover, we will analyse the trainings of a particular training plan, in order to ckeck if the training plan is focusing on the parameters seen before. The study of a particular training plan will require to find how some training indicators have been computed in relation to theoretical training's parameters.

The method we are going to use to achieve our purposes is Principal Component Analysis, since it provides the relation between the variables of a dataset and it allows us to reduce de dimension in order to have a better interpretation of it. For us, it will be usefull to know which are the most important parameters of a running training plan and how are them related.

## 1.4 Outline

In this report we are going to see how the parameters collected from many running trainings sessions will define how a training plan is needed to be planned, in the sense of focusing on training specific parameters. We expect that from all the parameters, only a few of them will be needed to be trained carefully, while the others will play less important roles.

The structure of the document will be the following:

In Chapter 2, we are going to see which kind of parameters can be found in a running training session. We will know what they describe and we will see all the technical aspects of running trainings in order to follow the hole report.

In Capter 3, the Principal Component Analysis will be briefly explained, defining and demonstrating the main theoretical results using algebraic and statistical tools.

In Chapter 4, a detailed evolution of the procedure from the initial questions to the final conclusions will be given. Moreover, we will see all the information related with the data used in the report: content, origin, date and quantity. Then, we are going to know how is this data adapted and treated.

In Chapter 5, results will be presented. We will use the method from Chapter 3 and we will see which information can we obtain from the analysis of running trainings.

Finally, in Chapter 6, we will discuss and present the conclusions of the results. We will end up giving some recommendations for training plans to increment their efficiency in terms of improving the velocity of the athlete during the training plan.

# Chapter 2

# Some running knowledge

There are a lot of kinds of running races. For example, if we classify them by the distance, we can find from 50 meters races to races with more than 100 kilometers. If we classify them by the elevation, we can find races with no elevation and races where the participants need to climb more than 2.000 meters to finish the race. Obviously, the training plan won't be the same if we are training for a 50 meters race or if we want to train for a 100 kilometers race as well as for elevation.

## 2.1 A running training session

### 2.1.1 Parameters

The parameters that can be collected from a running training session are:

1. **Distance**: Total amount of kilometers and meters run.
2. **Time**: Total amount of hours, minutes and seconds to complete the training session. If the runner stops, this time is not considered until he starts running again.
3. **Pace**: Relation between time and distance. It can be considered the average of all the training, the minimum value or the average of some interval of length. It is expressed as how much time will be needed to complete a kilometer. For example, if we run 5 kilometers in 20 minutes, we will say that the average pace has been 4 minutes per kilometer.
4. **Heart rate**: Relation of the number of beats of our heart and time. As in the case of Pace, we can consider the average, the maximum or the minimum of the heart rate during a training.
5. **Elevation gain**: Sum of every gain in elevation.

Notice that if we have the Distance and the Pace, we will know the Time by computing Distance times Pace. For this reason, we won't consider Time any more.

Another peculiarity is that velocity is not used as a parameter because we have already the Pace, which is the inverse of velocity. Between athletes, it is common to use Pace instead of velocity and we will use it too during all the report for convenience.

### 2.1.2 Types of running training

There are two different types of running trainings. On one hand we have the **Base runs**, that are those training session where the athlete starts running and keeps more or less the same velocity during the hole session. On the other hand, we have the **Intervals**, that are those trainings sessions where the athlete starts with a brief warm-up; then, for short distances, starts running as fast as he can and stops, when he has completed the distance, a short period of time. This procedure will be repeated a number of times and finally the athlete ends the training with a calm-down.

In some running training plans it can be observed that more than one session per day is being planned. So, it will be necessary to define a new variable to mark whether if it is the first time of the day that the athlete trains or not because the fatigue of the first session may affect the second one. Let us call it as **Post-activity**.

### 2.1.3 Theoretical trainings

A running training plan tells, for every session, what is the athlete supposed to do. We can think as if it is a prediction of all the trainings. These predictions are called **Theoretical trainings**. In order to compare the training done by the the athlete and the theoretical one, it is defined the **Intensity Factor (IF)** [3]. This coeficient is an indicator of how good has been a training in terms of effort. The higher is its value, the more effort will be done in the training.

These theoretical trainings have been calculated using the results of a stress test [8]. This test consists on running in a treatmill, starting with a really slow pace and going incrementing it every minute, until the limit of the athlete. This test provides, among other results, the values where the athlete effort changes. The following picture is an example of how the heart rate and the pace are classified in seven intervals depending on the values of the test:

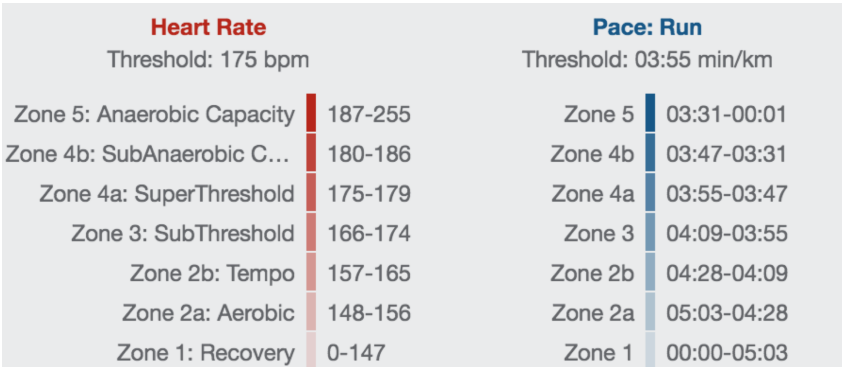



Figure 2.1: Example of zones of work of an athlete.

As we can see, we have all the values distributed in one of these zones. Moreover, every zone of the heart rate is related with the same level zone of the pace. There is also the **threshold value** for both parameters, which tells the specific value where the athlete effort changes from aerobic to anaerobic. Finally, it is important to say that this table is different for every person, because everyone has his/her own genetics and his/her own fitness level.

# Chapter 3

# Method of principal component analysis

## 3.1 Introduction

Principal Component Analysis (PCA) is a popular technique used in a lot of fields like image processing [2], driving simulations [10]... As we will see in the following chapters, we are going to use this method to determine which are the most important parameters to analyse in a running training plan. In other words, we will find the variables where we have more variation, and we will focus on studying how is their behaviour and which are their trends.

As we are going to see in this chapter, Principal Component Analysis' method is based on the reduction of the dimension of our dataset, preserving as much information as possible. Firstly, it defines a new orthonormal basis composed by linear combinations of our initial variables of the dataset, called Principal Components. Then, it keeps only the Principal Components with the largest variance, so we will have a new space with less dimensions that contains as much information as we want from the initial one. These Principal Components are where our dataset has no correlation and where most of the statistical information is being concentrated, so we are able to do a better interpretation of it and work easily with it due to the reduction of the dimensionality.

The most important results of the PCA method and the Singular Value Decomposition (SVD) [9] [11], will be given as an alternative of the diagonalization of the covariance matrices. The detailed explanation of this method can be found in the reference section [5] [7].

## 3.2 Principal Component Analysis

**Definition 3.2.1.** Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ be a $m$-dimensional variable ($\mathbf{x} \in \mathbb{R}^m$). Let $\mathbf{x}^1$, $\mathbf{x}^2$, ...,$\mathbf{x}^n$ be $n$ samples of our variable $\mathbf{x}$. Suppose that we have more samples than dimensions of the variable ($m \ll n$).Then, we can express the *sample data matrix* ($X$) as the $m \times n$ matrix:

$$X = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \mathbf{x}^n \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}$$

where $\mathbf{x}^i = (x_1^i, x_2^i, \ldots, x_m^i)$, $\forall i = 1, \ldots, n$.

### 3.2.1 Covariance matrix $C_1$

Once we have our sample data matrix defined, our purpose is to obtain an orthonormal basis of our dataset, preserving as much information as possible. This is, in fact, to find those directions(Principal Components) where $X$ has more variability, keeping orthogonality. For this, we need to define the *sample covariance matrix*:

**Definition 3.2.2.** Let $X$ be our *sample data matrix*, with $X \in \mathcal{M}_{m\times n}(\mathbb{R})$. The *sample covariance matrix* ($C_1$) of our variable $x$ is:

$$(C_1)_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j) \quad 1 \leq i, j \leq m \tag{3.1}$$

where $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)$ is the mean value for each variable, with $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_i^k$, $\forall i = 1, \ldots, m$.

Notice that if we have the m variables with mean equal to zero, this is $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m) = (0, 0, \ldots, 0)$, then the expression for $C_1$ turns:

$$C_1 = XX^T \tag{3.2}$$

We can assume, without loss of generality, that all of the $m$ variables have mean equal to zero. The reason is that if we have $n$ samples of a $m$-dimensional variable $\mathbf{x}$ with mean $\bar{\mathbf{x}}$, then we can define new $n$ samples subtracting the mean at every sample: $\mathbf{y}^i = \mathbf{x}^i - \bar{\mathbf{x}}$, $\forall i = 1, \ldots, n$. The factor $\frac{1}{n}$ is just for scaling and can be neglected. Up to this point, we will supose that the mean of the $m$ variables is zero and we will use the second definition for the covariance matrix $C_1$. It is important to see that the matrix $C_1$ has the following properties:

1. $C_1$ is a $\mathcal{M}_{m\times m}(\mathbb{R})$ squared matrix.
2. $C_1$ is symmetric matrix.
3. $C_1$ is a positive-semidefinite positive.

Now, due to $C_1$ is a symmetric and positive-semidefinite matrix, we are able to apply the *Spectral Theorem*:

**Theorem 3.2.3** (Spectral Theorem)**.** Let $A \in \mathcal{M}_{m\times m}(\mathbb{R})$ be a symmetric, positive-definite matrix of real values. Then

1. $A$ has $m$ real eigenvalues that satisfy $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$.
2. $A$ has $m$ eigenvectors $u_1, u_2, \ldots, u_m$ that satisfy $Au_i = \lambda_i u_i$, $\forall i = 1, \ldots, m$ and generates an orthonormal basis of $\mathbb{R}^m$.
3. $A$ can be expressed as $A = SDS^{-1} = SDS^T$, where $D = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ is a diagonal matrix with the eigenvalues of $A$ and $S = (u_1, u_2, \ldots, u_m)$ is an orthonormal matrix with the eigenvectors of $A$ in column.

Thank to the *Spectral Theorem*, our covariance matrix decomposes in:

$$C_1 = U\Lambda_1 U^T \tag{3.3}$$

where $U \in \mathcal{M}_{m\times m}(\mathbb{R})$ is the orthonormal matrix and $\Lambda_1 \in \mathcal{M}_{m\times m}(\mathbb{R})$ is the diagonal matrix resulting of diagonalizing $C_1$.

**Definition 3.2.4.** Let $\mathbf{z} = U^T\mathbf{x}$ be the expression of the variable $\mathbf{x}$ in the new orthonormal basis. Then, we can define $Z = U^T X$ as *the projected sample matrix* of $\mathbf{x}$ (matrix of samples of $\mathbf{z}$).

**Proposition 3.2.5.** The covariance matrix of the *the projected sample matrix* $Z$ is the diagonal matrix $\Lambda_1$.

*Proof.* Taking the definition of the covariance matrix for $X$ (2):

$$ZZ^T = (U^T X)(U^T X)^T = (U^T X)(X^T U) = U^T X X^T U = U^T C_1 U =$$
$$= U^T (U\Lambda_1 U^T)U = \Lambda_1$$

as $U$ is orthonormal and $C_1 = U\Lambda_1 U^T$ □

### 3.2.2 Reducing the dimension

We have seen that the covariance matrix of $Z$ is $\Lambda_1$, which is diagonal. This means that the $m$ components of $\mathbf{z}$ are fully uncorrelated (and linearly independent). This is, in fact, that the variance of every variable of $\mathbf{x}$ is described by the corresponding component of $\mathbf{z}$. The following step is to reduce the dimension, preserving as much information as possible. So, we are going to use the trace and the eigenvalues of both covariance matrices $C_1$ and $\Lambda_1$ because they will tell us how much dispersion accumulates each component of both $X$ and $Z$.

It is important to see that the trace and the eigenvalues of $C_1$ and $\Lambda_1$ are the same due to trace and eigenvalues are invariant under change of basis:

**Proposition 3.2.6.** The trace of $C_1$ is the same as the trace of $\Lambda_1$ and their value is $\sum_{i=1}^{m} \lambda_i$.

*Proof.* As we have seen in the previous section, $C_1 = U\Lambda_1 U^T$. Then, if we want to compute the trace of $C_1$:

$$tr(C_1) = tr(U\Lambda_1 U^T) = tr(\Lambda_1 U U^T) = tr(\Lambda_1) = \sum_{i=1}^{m} \lambda_i$$

where we have used the property of $tr(AB) = tr(BA)$ with $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ and $B \in \mathcal{M}_{n\times m}(\mathbb{R})$. □

**Definition 3.2.7.** Let $\epsilon \in (0, 1)$ be the *information loss parameter*. This parameter will control the maximum of information lost when we reduce the dimension. Notice that, the lower the value of $\epsilon$ is, the less the information is lost. So, we will take $\epsilon \ll 1$.

**Definition 3.2.8.** Let $r = 1, \ldots, m$ be the *reduced number of dimensions*. This is the minimum number of dimensions of the reduced space needed to perserve as much information as we want.

The relation between the *information loss parameter* ($\epsilon$) and the *reduced number of dimensions* ($r$)is the following:

$$\sum_{i=1}^{r} \lambda_i \geq (1 - \epsilon) \sum_{i=1}^{m} \lambda_i \tag{3.4}$$

This reduction of the dimension neglect the $m - r$ lower eigenvalues and, consequently, their associated eigenvectors. But these eliminated eigenvalues are the ones that tell us less information because they are related to the components with lower variance. On the other hand, we are keeping the largest eigenvalues so, the directions that contain more statistical information of our dataset.

Once we define how much information we allow to be lost (give a value for $\epsilon$), we obtain a value for $r$ and we are able to define the *new reduced basis* and the *reduced projected sample data matrix*:

**Definition 3.2.9.** We define the *new reduced basis* $U^* \in \mathcal{M}_{m\times r}(\mathbb{R})$ as the first $r$ eigenvectors of $U$.

**Definition 3.2.10.** We define the *reduced projected sample data matrix* $Z^* \in \mathcal{M}_{r\times n}(\mathbb{R})$ as the first $r$ dimensions of the *projected sample matrix* $Z$. $Z^*$ is obtained from the *new reduced basis* $U^*$ and the *sample data matrix* $X$ by computing the product:

$$Z^* = (U^*)^T X \tag{3.5}$$

Summarising, we have found a new space where the variables are not correlated, with less dimensions than the original one and that contains as much information as we want. For this reason, working in this new space will allow us to do better interpretations of our data and have results easier.

Finally, it is important to know that we can return to the initial space once we have used the reduced-dimension space by the backward mapping:

$$\hat{X} = U^* Z^* \tag{3.6}$$

In order to compare our *sample data matrix* $X$ with $\hat{X}$ and we see that the error produced by the reduction of dimensionality is related to the value of the tolerance $\epsilon$. The error increases as the value of $\epsilon$ increases. So, we can conclude that we can approximate $X$ by $\hat{X}$:

$$X = UZ \approx U^* Z^* \tag{3.7}$$

### 3.2.3 Covariance matrix $C_2$

Up to now we have been considering that we have more samples of a variable $\mathbf{x}$ than dimensions of this variable ($m \ll n$). In the case that we are in the opposite situation ($n \ll m$), it is useful to consider samples as elements of an $n$-dimensional space. This choise will avoid us to compute the covariance matrix $C_1$, which will be more expensive in therms of computational cost. for this, we define the new covariance matrix:

**Definition 3.2.11.** We define the *sample covariance matrix* $C_2$ as:

$$(C_2)_{ij} = \frac{1}{m} \sum_{k=1}^{m} (x_k^i - \bar{x}^i)(x_k^j - \bar{x}^j) \qquad \forall i, j = 1, \ldots, n \tag{3.8}$$

where $\bar{x}^i = \frac{1}{m} \sum_{l=1}^{m} x_l^i$ is the mean value of the sample $\mathbf{x}^i$.

As we have seen for the *sample covariance matrix* $C_1$, if $\bar{x}^i = 0 \; \forall i = 1, \ldots, n$, then we can express the *sample covariance matrix* $C_2$ as:

$$C_2 = X^T X \tag{3.9}$$

In this case we can also assume, without loss of generality, that the mean value for the sample is 0 by the same argument as for the covariance matrix $C_1$. The properties of $C_2$ are the same as the properties of $C_1$, but in this case $C_2$ is a $\mathcal{M}_{n \times n}(\mathbb{R})$ matrix. So, the covariance matrix satisfies the *Spectral Theorem* conditions and we can decompose it in:

$$C_2 = V \Lambda_2 V^T \tag{3.10}$$

where $V \in \mathcal{M}_{n \times n}(\mathbb{R})$ is a orthogonal matrix and $\Lambda_2 \in \mathcal{M}_{n \times n}(\mathbb{R})$ is a diagonal matrix with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$.

Now, we are able to compute the *projected sample matrix* $Z$ in therms of $V$ and $X$.

$$Z = V^T(X^T X) \tag{3.11}$$

*Proof.* The equivalence of the two definitions of $Z$ is going to be seen in section 3.3. □

Finally, the dimensionality reduction is computed in the same way as for the *covariance matrix* $C_1$, arriving to the same results.

### 3.2.4 The equivalence of diagonalizing $C_1$ and $C_2$

The only thing we need to see is that we will get the same result if we compute $C_1$ or if we compute $C_2$. Suppose that we are in the case when our *sample data matrix* $X$ has more samples of our variable than dimensions of the variable ($m \ll n$). The other case is done in the same way permuting $C_1$ and $C_2$.

We have seen how to compute both covariance matrices $C_1$ and $C_2$ and how to diagonalize them. Then applying those definitions in our case and diagonalizing them, we have that

$$C_1 u^i = \lambda_i u^i \qquad \forall i = 1, \dots, m \tag{3.12}$$

and

$$C_2 v^i = \lambda_i v^i \qquad \forall i = 1, \dots, m \tag{3.13}$$

the other $n - m$ eigenvalues are 0 and their associated eigenvectors describe the kernel space of $C_2$. This means that they do not have information of our dataset, so when we reduce de dimension we are not taking them account. For this reason, due to we have the same non-zero eigenvalues for $C_1$ and $C_2$, the dimensionality reduction is the same for both matrices.

### 3.2.5 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is an alternative method to have a decomposition of the *covariance matrix* $C_1$ and $C_2$ because it avoids diagonalization [11]. We will only apply the method once to the *sample data matrix* and we will end having both expressions for the matrices. So, if we take the (SVD) for the *sample data matrix* $X$:

$$X = U\Sigma V^T \tag{3.14}$$

where $U \in \mathcal{M}_{m\times m}(\mathbb{R})$ and $V \in \mathcal{M}_{n\times n}(\mathbb{R})$ are orthogonal matrices and $\Sigma \in \mathcal{M}_{m\times n}(\mathbb{R})$ is a diagonal matrix with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq 0$.

Using this expression of our matrix $X$, we can obtain the decomposition of the covariance matrix $C_1$:

$$C_1 = XX^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T = U\Lambda_1 U^T \tag{3.15}$$

where $\Lambda_1 = \Sigma\Sigma^T$. See that $\Lambda_1 \in \mathcal{M}_{m\times m}(\mathbb{R})$ diagonal with values $\lambda_i = \sigma_i^2, \forall i = 1, \ldots, m$.

We can also consider $C_2 = X^T X$ and, as before, $C_2$ can be expressed as a result of the (SVD) of $X$:

$$C_2 = X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T = V\Lambda_2 V^T \tag{3.16}$$

where $\Lambda_2 = \Sigma^T\Sigma$. See that $\Lambda_2 \in \mathcal{M}_{n\times n}(\mathbb{R})$ diagonal with values $\lambda_i = \sigma_i^2, \forall i = 1, \ldots, m$ and $\lambda_j = 0, \forall j = m+1, \ldots, n$.

To end up with this section, we have seen that with only the decomposition of the *sample data matrix* $X$ using the (SVD), we can obtain both covariance matrices $C_1$ and $C_2$ and the change of basis that transform them into diagonal matrices. At the same time, we can see that the two definitions of the *projected sample matrix* $Z$ are equivalent and generate the same vectorial space:

*Proof.* Recalling both definitions and using the (SVD) of $X$:

$$Z = U^T X = U^T U\Sigma V^T = \Sigma V^T$$
$$Z = V^T(X^T X) = V^T(V\Sigma^T U^T U\Sigma V^T) = V^T(V\Sigma^T\Sigma V^T) = \Sigma^T\Sigma V^T$$

And we observe that $< \Sigma V^T >=< \Sigma^T\Sigma V^T >$ due to $\Sigma^T\Sigma \in \mathcal{M}_{n\times n}(\mathbb{R})$ and $\Sigma \in \mathcal{M}_{m\times n}(\mathbb{R})$ are diagonal. $\square$

# Chapter 4

# Methodology

As we have seen in previous chapters, in this project we are going to deal with running trainings, using the information that they provide to analyse, interpret and extract some final concluisons about them. We are going to apply the Principal Component Analysis method to find out which are the most important parameters by looking into their variability in order to know where we need to focus on our trainings.

The data we are going to work with has been obtained from a database that collect all the performs of a running smartwatch with GPS and a heart rate sensor integrated: Garmin Forerunner 935. The account where the watch is linked is my own Garmin Connect profile. In this webpage, we can find a register of my daily workouts and many personal phisical status such as body weight, fitness mark, or the threshold value for the pace (4 minutes 20 seconds) and the heart rate (175 bpm). Inside the workouts, we can see a summary of our training with the route, whether contidions and all of the parameters achieved like distance, average pace, average heart rate, elevation gain, ... There is also a segmentation of the distance in kilometers and their respective paces and some graphics showing the elevation gain, the pace and the heart rate through the time.

The period of time of our running trainings is from 1st of January of 2019 to 31st of December of 2019. During all this time, I have completed a total of 80 running trainings, focused on 5 and 10 kilometers races. It is important to mention that these workouts have been combined with others swimming and cycling trainings, which have not been considered. The most common parameters of a running training session are distance, average pace, average heart rate, elevation gain, Intensity Factor (IF), Intervals and Post-Activity.

Now, we are in conditions to present the variables we will use to describe a training. For this, we define $\mathbf{x}$ as:

$$\mathbf{x} = \begin{bmatrix} \text{Distance } (km) \\ \text{Average pace } (min/km) \\ \text{Average heart rate } (bpm) \\ \text{Elevation gain } (m) \end{bmatrix}$$

And we will leave IF, Post-Activity and Intervals as independent variables. It is important to notice that the variables are not in the same units. Due to we cannot compare and analyse variables with different units, we need to put them in the same ones. We will choose:

1. Kilometers for distance.
2. Seconds for time.
3. Beats per minute for the heart rate.

Although we can see that we are using minutes for the heart rate, beats per minute (bpm) is the standard unit for heart rate measurements. For this reason we will take bpm as heart rate unit instead of beats per second.

Once we have put all the variables in the same units (notice that Post-Activity, Intervals and IF are dimensionless), the next step is to define our *sample data matrix* as:

$$X = \begin{bmatrix} \mathbf{x}_1 \dots \mathbf{x}_{80} \end{bmatrix} = \begin{bmatrix} \text{Distance}_1 & \dots & \text{Distance}_{80} \\ \text{Average pace}_1 & \dots & \text{Average pace}_{80} \\ \text{Average heart rate}_1 & \dots & \text{Average heart rate}_{80} \\ \text{Elevation gain}_1 & \dots & \text{Elevation gain}_{80} \end{bmatrix}$$

After that, we need to center our samples. This means that we need to calculate the mean value of every variable $\mathbf{x}^j$ :

$$\bar{\mathbf{x}}^j = \frac{1}{80} \sum_{i=1}^{80} \mathbf{x}_i^j \qquad \forall j = 1 \dots 4 \tag{4.1}$$

where $\mathbf{x}_i^j$ is the $j$-th variable of the $i$-th training. Subtracting this mean value for every variable in all the samples, we obtain a zero mean matrix $X_c$:

$$X_c = \begin{bmatrix} \mathbf{x}_1 - \bar{x} \dots \mathbf{x}_{80} - \bar{x} \end{bmatrix}$$

where

$$\bar{x} = \begin{bmatrix} \bar{x}^1 \\ \bar{x}^2 \\ \bar{x}^3 \\ \bar{x}^4 \end{bmatrix}$$

It will be useful to keep the non-centered matrix $X$ to observe som data behaviours and intuitive relations, so we will name the centered one as $X_c$. For example, we will analyse the relation between the variables Average pace and Average heart rate by aproximating them using linear least squares [1]. The aim of this method is to approximate the variables by a first degree polynomial minimizing the sum of the squares of the residuals of the samples to the polynomial.

**Proposition 4.0.1** (Computation of the coefficients for linear least squares approximation)**.** Supose we have $n$ samples of a 2-dimensional variable $P = (r, s)$. We want to find $c_1$ and $c_2$ such that

$$s = c_1 r + c_2 \tag{4.2}$$

is the best approximation of the samples that minimizes

$$SUM = \sum_{i=1}^{n} (s_i - (c_1 r_i + c_2))^2 \tag{4.3}$$

*Proof.* The last equation in vectorial form is:

$$SUM = s^T s - 2c_1 s^T r - 2c_2 n\bar{s} + c_1^2 r^T r + 2c_1 c_2 n\bar{r} + nc_2^2 \tag{4.4}$$

where $\bar{s}$ and $\bar{r}$ are the mean value of s and r. To minimize, taking the derivative of $c_1$ and $c_2$ and equalizing to 0, we obtain:

$$\frac{\partial SUM}{\partial c_1} = -2s^T r + 2c_1 r^T r + 2nc_2\bar{r} = 0 \tag{4.5}$$

$$\frac{\partial SUM}{\partial c_2} = -2n\bar{s} + 2nc_1\bar{r} + 2nc_2 = 0 \tag{4.6}$$

Finally, solving this linear system of equations, we end up with the following expression for the two coefficients:

$$c_1 = \frac{s^T r - n\bar{r}\bar{s}}{r^T r - n\bar{r}^2} \tag{4.7}$$

$$c_2 = \bar{s} - c_1\bar{r} \tag{4.8}$$

□

Once we will have studied some data particularities, we will start with the Principal Component Analysis (PCA). For this, we will use the centered *sample data matrix* $X_c$, we will compute the matix $C_1$ and we will diagonalize it applying the Singular Value Decomposition. Due to $C_1$ is a squared symmetric matrix, we will obtain the following decomposition of $C_1$:

$$C_1 = U\Sigma U^T \tag{4.9}$$

where $U \in \mathcal{M}_{4\times 4}(\mathbb{R})$ orthogonal and $\Sigma \in \mathcal{M}_{4\times 4}(\mathbb{R})$ diagonal.

Then, we will reduce de dimension and we will analyse the results in the new reduced space. However, we will observe that they won't show enough relevant information. So, we will decide to add more specific parameters to our matrix such as the minimum pace and the maximum heart rate of every training, creating the new matrix $\tilde{X}$. These two new variables will provide more detailed information of the pace and the heart rate of the training, knowing their extreme values.

As before, we will need to adapt both variables by changing their units and subtracting their mean for every sample, resulting the matrix $\tilde{X}_c$. These two new variables will bring to our matrix new statistical information that, thanks to PCA again, we will see that it provides new more useful results.

To conclude with the result's chapter, we will study our particular training plan, analysing those important parameters and checking if it helps the athlete to improve them. For this, we will need to have the length of the interval trainings to compare the ones that have the same length. So, we will plot the evolution of the parameter every interval training during the hole year. Moreover, we will try to find how is IF computed in order to determine if it is a good indicator to describe every training. As a first thing to try, we will represent the two first components of the reduced space with the IF and we will look for some relation. Then, due to we want to express IF using the initial data, we will turn back to the initial space and we will approximate IF using linear least squares, obtaining a formula to calculate IF from the information of a training.

Finally, we will discuss on the results and we will end up with conclusions and recommendations for running training plans.

# Chapter 5

# Results

## 5.1 Initial data

In this chapter we are going to see the results of our project and their analysis. As we have seen in the previous chapter, we are studying running trainings, which are described by the following variables:

$$\mathbf{x} = \begin{bmatrix} \text{Distance} \\ \text{Average pace} \\ \text{Average heart rate} \\ \text{Elevation gain} \end{bmatrix}$$

In order to know some things of our *sample data matrix* $X$, it may be convenient to plot some of the parameters. For example, intuitively the relation between average pace and average heart rate needs to be decreasing. To see this relation, we will compute the linear approximation using least squares of the samples and we will check if its slope is negative. So, we want to find $A$ and $B$ such that:

$$\text{Average heart rate} = A \cdot \text{Average pace} + B \tag{5.1}$$

is the best linear approximation of the samples using a first degree polynomial. As we can see in Proposition 4.0.1, if we denote

$y = \text{Average\_heart\_rate}$

$t = \text{Average\_pace}$

The computation of $A$ and $B$ using vectorial calculus turns to:

$$A = \frac{(t)^T * (y) - 80 * \bar{t} * \bar{y}}{(t)^T * (t) - 80 * \bar{t} * \bar{t}} \tag{5.2}$$

$$B = \bar{y} - A * \bar{t} \tag{5.3}$$

where $\bar{t}$ and $\bar{y}$ are the mean value of each variable. The result of these computations is $A = -4.58{\cdot}10^{-2}$ and $B = 1.80{\cdot}10^{2}$.

As we can see, $A$ is negative, so the slope of the linear approximation is negative and the relation of Average heart rate and Average pace is decreasing, as we thought.

In the following picture we can observe this relation: the lower the average pace is, the higher the average heart rate is.

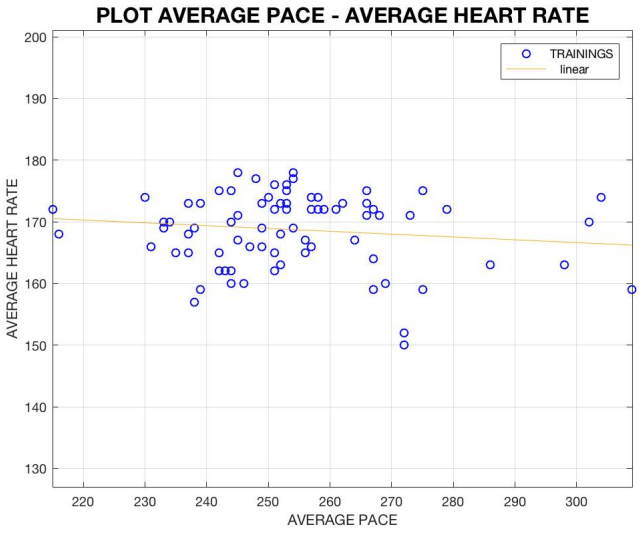



Figure 5.1: Plot of Average pace vs Average heart rate

Once we have seen some particularities of our initial data, the following step is to do a change of basis where our data is uncorrelated. For this, we take the centered sample data matrix $X_c$ and we apply the Principal Component Analysis method.

### 5.1.1 Study of the initial data

First, we start computing the covariance matrix $C_1$ using the Equation 3.2, due to our matrix $X_c$ has mean equal to 0 for all of the four variables.

The results obtained from the orthonormal decomposition of $C_1$ are the following eigenvalues (with the amount of information) and their associated eigenvectors:

| **Information of the eigenvalues** | | |
|---|---|---|
| Eigenvalue | Eigenvector | Percentage |
| $\lambda_1 = 2.42{\cdot}10^4$ | $\mathrm{U}(:,1) = \begin{pmatrix} 0.00 \\ 0.99 \\ -0.05 \\ 0.00 \end{pmatrix}$ | 88.3 % |
| $\lambda_2 = 2.86{\cdot}10^3$ | $\mathrm{U}(:,2) = \begin{pmatrix} -0.11 \\ -0.05 \\ -0.99 \\ 0.00 \end{pmatrix}$ | 10.4 % |
| $\lambda_3 = 3.57{\cdot}10^2$ | $\mathrm{U}(:,3) = \begin{pmatrix} 0.99 \\ 0.00 \\ -0.11 \\ 0.00 \end{pmatrix}$ | 1.3 % |
| $\lambda_4 = 0.70$ | $\mathrm{U}(:,4) = \begin{pmatrix} 0.00 \\ 0.00 \\ 0.00 \\ 1.00 \end{pmatrix}$ | 0.002 % |

Table 5.1: Information obtained from the change of basis of $C_1$

It is important to notice that the matrix $U$ that contains the eigenvectors of $C_1$ is quite similar to the $Id$ matrix. This means that our initial data is almost already uncorrelated and all of the parametres we are working with are independent. We can also see that the most relevant parameter is the Average pace, followed by the average heart rate, the distance and the elevation gain.

This information shows some interesting things that are important to comment. First of all, due to the average pace is independent from the distance, we can say that it doesn't matter which is the distance of the race we are training for, the training plan will be focused on improving the pace for 5 or 10 kilometres races so, it won't be necessary to worry about the distance of the training. Secondly, due to the average pace is independent from the average heart rate, that there is a value for the pace that makes the heart rate to loss some dependcy on the pace instead of phisical and genethic properties. This value could be the threshold value explained in Chapter 2.

Next step is to choose an appropiate value for $\epsilon$, which will say how many Principal Components do we need to take in order to preserve as much information as we want. Using $\epsilon = 0.015$, we see that only the two first Principal Components are needed. This is, in fact, to take $r = 2$ for the dimensionality reduction.

Defining

$$Z_1^* = U(:,1)^T X_c \tag{5.4}$$

$$Z_2^* = U(:,2)^T X_c \tag{5.5}$$

where $U(:,i)$ is the eigenvector asociated to the eigenvalue $\lambda_i \ \forall i = 1 \dots r$, we have the new reduced basis. In order to introduce the Interval and Post-Activity parameters as a postprocess, we start colouring and shaping the samples of the reduced space using both parameters and we obtain the figure below. For convenience, we denote $Z_i^* = Zsti \ \forall i = 1 \dots r$ in the plots from now on until the end of this chapter. As we see in the legend, an Interval training corresponds to the blue mark and a Post-Activity training corresponds to a * mark.

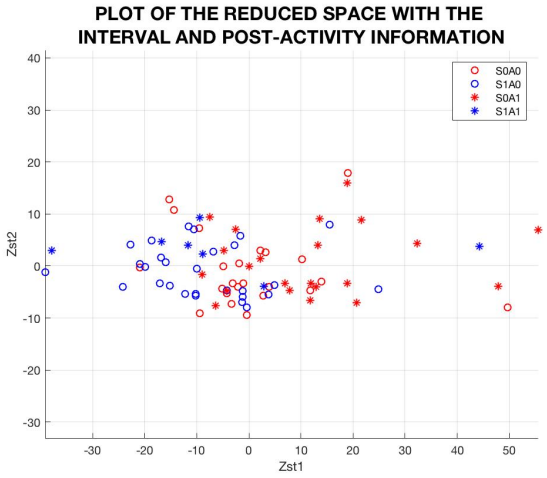



Figure 5.2: Plot of the reduced space with the interval and post-activity information.

We can see that we have the nearly the same plot (with a $Z_2^*$ symmetry) as the average pace vs average heart rate one seen before (one of the results seen from the diagonalization of the covariance matrix $C_1$ is that our parameters are uncorrelated). Notice that there is a tendency to have the intervals workouts on the left and the base runs on the right, but not in a clear way. It is important to remember that, in an interval session, sometimes the stopwatch is not paused between intervals and this period of time is affecting the final average pace. Also, the warm-up and the calm-down are modifying the final value due to they are usually done slowlier.

For this reason we can say that using this four parameters, a running training is not being well described. To solve this inconvenience, we add two more speific parameters to our matrix in order to obtain more detailed information from every training.

## 5.2 Adding more specific parameters

Up to now we have seen some interesting results like the equivalence to train for a 5 quilometres or a 10 quilometres race or the non dependency on the heart rate of the pace when we are running in slower threshold paces. Although they are great results, we want to check if we can describe our trainings more precisely, in order to know where do we need to focus on while training. For this reason, we redefine our variable $x$ by adding two more parameters, which are the maximum heart rate and the minimum pace of every training. Then it seems:

$$\mathbf{x} = \begin{bmatrix} \text{Distance} \\ \text{Average pace} \\ \text{Average heart rate} \\ \text{Elevation gain} \\ \text{Maximum heart rate} \\ \text{Minimum pace} \end{bmatrix} \tag{5.6}$$

Thanks to these two new parameters, we are going to control the variation of the pace and the heart rate during a training. We would expect that for base runs the variation of pace is lower than for interval trainings and the same for the heart rate.

### 5.2.1 Study of the new database

In this section we are going to repeat all the process to uncorrelate the variables by doing a change of basis and taking those Principal Components such that the most part of the data information is being preserved.

Now, the results of the diagonalization are:

| **Information of the eigenvalues** | | |
|---|---|---|
| Eigenvalue | Eigenvector | Percentage |
| $\lambda_1 = 7.20{\cdot}10^4$ | $\mathrm{U}(:,1) = \begin{pmatrix} -0.01 \\ -0.28 \\ 0.00 \\ 0.00 \\ 0.04 \\ -0.96 \end{pmatrix}$ | 74.8 % |
| $\lambda_2 = 2.00{\cdot}10^4$ | $\mathrm{U}(:,2) = \begin{pmatrix} 0.02 \\ -0.96 \\ 0.06 \\ 0.00 \\ -0.06 \\ 0.28 \end{pmatrix}$ | 20.7 % |
| $\lambda_3 = 3.42{\cdot}10^3$ | $\mathrm{U}(:,3) = \begin{pmatrix} -0.09 \\ -0.03 \\ -0.89 \\ 0.00 \\ -0.45 \\ -0.01 \end{pmatrix}$ | 3.5 % |
| $\lambda_4 = 6.00{\cdot}10^2$ | $\mathrm{U}(:,4) = \begin{pmatrix} 0.17 \\ 0.07 \\ 0.43 \\ 0.00 \\ -0.88 \\ -0.06 \end{pmatrix}$ | 0.6 % |
| $\lambda_5 = 3.43{\cdot}10^2$ | $\mathrm{U}(:,5) = \begin{pmatrix} -0.98 \\ 0.00 \\ 0.16 \\ 0.00 \\ -0.12 \\ 0.00 \end{pmatrix}$ | 0.4 % |
| $\lambda_6 = 6.79$ | $\mathrm{U}(:,6) = \begin{pmatrix} 0.00 \\ 0.00 \\ 0.00 \\ -1.00 \\ 0.00 \\ 0.00 \end{pmatrix}$ | 0.0007 % |

Table 5.2: Information obtained from the diagonalization of the covariance matrix.

One of the new information that we observe is that the most important parameter is not the average pace as we saw before. In this case, it is the minimum pace. We continue seeing that the parameters present uncorrelation, even though less than before and the third and fourth eigenvectors are linear combinations of maximum heart rate and average heart rate. To reduce the dimension, we take the same $\epsilon = 0.015$ and we see that $r = 3$ is needed.

Now it is time to use the parameters Interval and Post-Activity in order to see any classification of the samples. The result of using them is the following image:

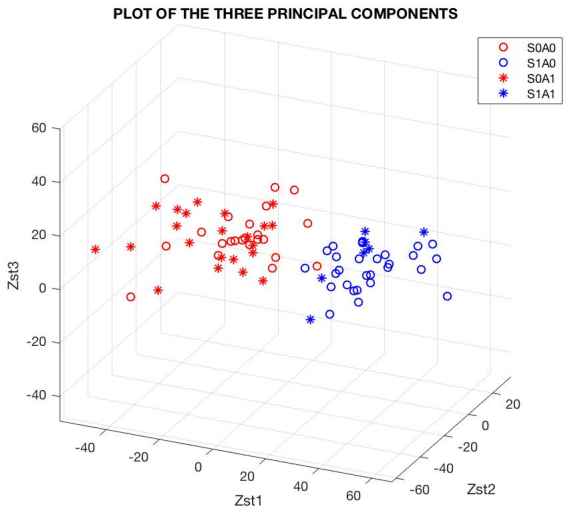



Figure 5.3: Plot of the three Principal Components.

We can see that the parameter Interval is really classified, due to we have base runs trainings on the left and intervals trainings on the right. This result gives us the type of run where we the minimum pace is trained, which is doing intervals sessions.

## 5.3 Analysis of our training plan

In the previous section we have determined that the most important parameter is the minimum pace of a training. This parameter is trained by doing many intervals sessions. So, we want to check if our training plan is performing as espected, allowing us to decrease the minimum pace.

To do so, we need to classify our intervals sessions. The reason is that the minimum pace while doing short intervals is not the same for every length of intervals. Observing them, we see that there are 4 different lengths of intervals: 300, 400, 1.000 and 2.000 meters.

The following pictures show, for every length of interval, the progression of the minimum pace every time we do an interval training of every length:

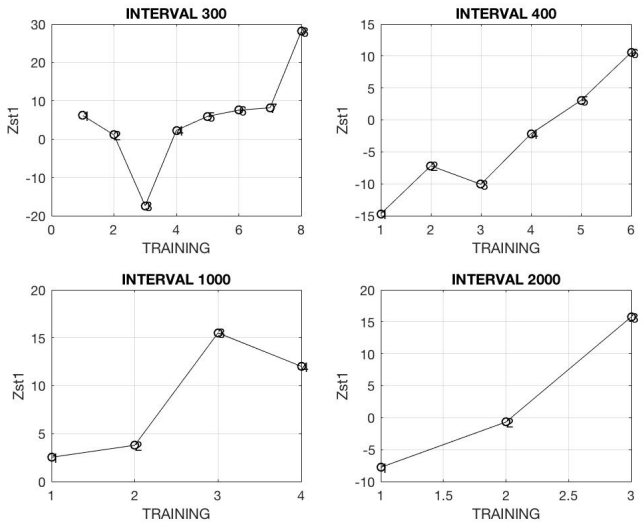



Figure 5.4: Evolution of the minimum pace for every type of interval. a) 300 meters. b) 400 meters. c) 1000 meters. d) 2000 meters.

For every length of interval we see that every time we do the same training, we reach lower and lower paces. Although it the graphics they are all increasing, we need to remember that in Table 5.2 the sign of the coordinate of the minimum pace of the eigenvector is negative, so in this case decreasing the minimum pace is seeing an increment of the values. This result brings us to confirm that our training plan is helping us to be faster.

### 5.3.1 Intensity Factor (IF)

As a final analysis of our training plan, we want to observe the parameter IF and try to see how is calculated. This information will help us to determine if IF is a good indicator of how good has been a training compared with the theoretical running training expected. As a first idea, we start plotting our Principal Components against IF and, for the second one, we see the following graphic:

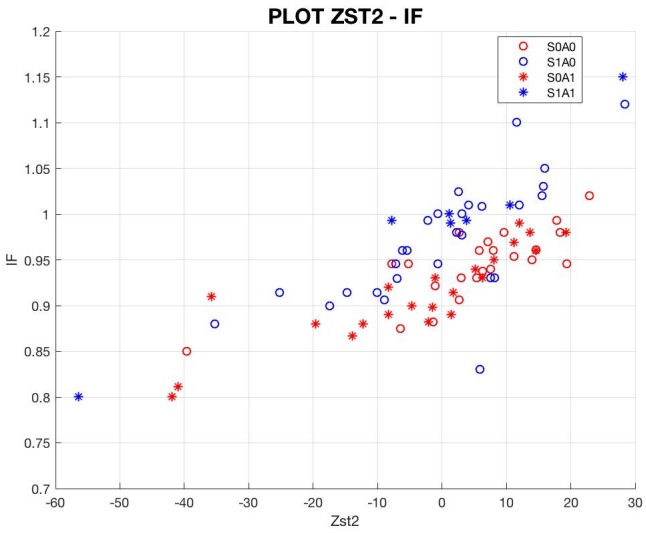



Figure 5.5: Plot of $Z_2^*$ vs IF.

Due to the second Principal Component is essentially the average pace with a few influence of the minimum pace, we deduce that IF is a value defined by the average pace. So, if we plot the Average pace versus IF, we obtain:

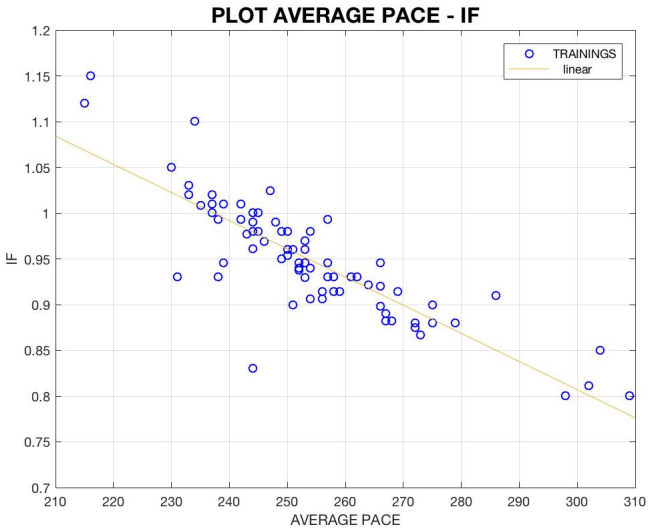



Figure 5.6: Plot of Average pace vs IF.

We can see that there is a linear behaviour. As in the first part of the chapter, we are going to approximate the samples in order to find a first degree polynomial that fits well our data.

Defining AP as Average pace, we want to find $\alpha$ and $\beta$ such that:

$$IF = \alpha * AP + \beta \tag{5.7}$$

Calculating this coefficients, we see that

$$\alpha = \frac{(AP)^T(IF) - 80(\bar{AP})(\bar{IF})}{(AP)^T(AP) - 80\bar{AP}^2} \tag{5.8}$$

$$\beta = \bar{IF} - \alpha\bar{AP} \tag{5.9}$$

where $\bar{AP}$ and $\bar{IF}$ are the mean value of each variable. So, the values are $\alpha = -3.01.10^{-2}$ and $\beta = 1.73$.

Thinking a little bit, we see that $\alpha \approx -\frac{1}{260}$. This value is the threshold value for the pace in seconds per kilometer units.

Summarising, we have found that IF formula is the following one:

$$\text{IF} \approx -\frac{\text{Average pace}}{\text{Threshold pace value}} + 1.73 \tag{5.10}$$

So, we can say that IF is just described by the average pace of the training.

# Chapter 6

# Discussion and conclusions

In this last Chapter we will present the conclusions, the discussion of the results seen in Chapter 5 and we will expose the porposals and needs for an efficient training plan. We will also comment on some points of our running training plan and their parameters as an example to follow in the case of checking other training plans.

1. **Training plans need to be focused on improving the minimum pace**. The main objetive of all training plans must be to prepare the athlete to be faster. This means that the athlete must be able to run in lower paces during more time. Thanks to Principal Component Analysis, we have seen that the most important parameter of a training is the minimum pace. The importance of improving this parameter is related with the comfortability of the athlete running in lower paces, letting him/her to spend less time to finish a race.

2. **Races of 5 and 10 kilometers have the same training plan**. One of the observations of the eigenvalues of the diagonalization of the covariance matrix has been that Distance does not have much relevance in a training as a parameter. This tells us, due to the training plan is prepared to train short distance races, that the runner does not need to worry about Distance while training for 5 or 10 kilometers races.

3. **Heart rate loses some pace dependance for low values of pace** This conclusion comes from the fact that we have seen that Average heart rate and Average pace are independent.

4. **Having a better description of a training helps to do a better analysis**. Using the most common parameters is not enough to describe completely a training. Most of the running trainings have different parts where the athlete needs to work harder or easier, so the average pace is not the best reference to tell how good has been a training session.

5. **Intensity Factor is described just by the average pace**. Although it seemed to be a sophisticated equation, it has been a relation of the threshold pace value and the average pace done in the training. Due to we have seen that the average pace is not a good indicator to describe a training, we suggest to change its formula by using also the minimum pace.

6. **Our training plan helps us to be faster**. The analysis of our training plan has shown that we have improved our minimum pace while following the training plan, as we can see in the interval trainings.

On a personal point of view, this project has shown to me an example of how mathematics can be used toghether with any field. It has been a great experience to apply data analysis with my hobbies and see how it gives such useful results that I will use for sure in my further trainings and races.

In conclusion, athletes will be required to work with mathematics if they want to improve and be the best. Up to now, the contribution of Big Data in sport has been extremly huge and this is just the beggining. For sure, we will see lots of new world records in the future.

# References

In this chapter we can find all the references used in this document:

[1] AUBANELL, Anton; BENSENY, Antoni; DELSHAMS, Amadeu. Útiles básicos de cálculo numérico. Servei de Publicacions de la Universitat Autònoma de Barcelona, 1993.

[2] CLAUSEN, Clifford; WECHSLER, Harry. Color image compression using PCA and backpropagation learning. Pattern Recognition, 2000, vol. 33, no 9, p. 1555-1560.

[3] COGGAN, A., 2020. Normalized Power, Intensity Factor And Training Stress Score. [online] Trainingpeaks.com. Available at: https://www.trainingpeaks.com/blog/normalized-power-intensity-factor-training-stress/ [Accessed 21 March 2020].

[4] Eliud Kipchoge Breaks Two-Hour Marathon Barrier, 2020. Nytimes.com [online] [Accessed 15 June 2020].

[5] GARCÍA-GONZÁLEZ, Alberto, et al. A kernel Principal Component Analysis (kPCA) digest with a new backward mapping (pre-image reconstruction) strategy. arXiv preprint arXiv:2001.01958, 2020.

[6] HICHENS, LIZ, 2018, Patrick Lange Goes Sub-8 Hours for First Time in Ironman World Championship History – Triathlete. Triathlete [online]. 2018. Available from: https://www.triathlete.com/events/patrick-lange-goes-sub-8- hours-for-first-time-in-ironman-world-championship-history/ [Accessed 14 June 2020].

[7] Jolliffe, I. T. Principal component analysis. Second Edition. Springer Series in Statistics, New York, 2002.

[8] KARVONEN, Juha; VUORIMAA, Timo. Heart rate and exercise intensity during sports activities. Sports medicine, 1988, vol. 5, no 5, p. 303-311.

[9] LÁZARO, J. Tomás; OLLÉ, Merce; PACHA, Juan R. Calcul Numeric. Manual de Practiques Facultat de Matematiques i Estadıstica UPC.

[10] LI, Mu; FU, Jia-Wei; LU, Bao-Liang. Estimating vigilance in driving simulation using probabilistic PCA. En 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2008. p. 5000-5003.

[11] M. Navas and C. Ordonez Efficient computation of PCA with SVD in SQL. Proceedings of the 2nd Workshop on Data Mining Using Matrices and Tensors, 2009.